Journal of Structural Biology 167 (2009) 136–144

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

Fold prediction of VP24 protein of Ebola and Marburg viruses using *de novo* fragment assembly

Michael S. Lee [a,b,*], Frank J. Lebeda [b], Mark A. Olson [b]

[a] Computational Sciences and Engineering Branch, US Army Research Laboratory, Aberdeen, MD 21005, United States
[b] Department of Cell Biology and Biochemistry, US Army Medical Research Institute of Infectious Diseases, Frederick, MD 21702, United States

ARTICLE INFO

Article history:
Received 13 January 2009
Received in revised form 4 May 2009
Accepted 7 May 2009
Available online 15 May 2009

Keywords:
De novo fragment assembly
Rosetta
Fold recognition
Ebola
Marburg
Hemorrhagic fever virus
VP24
Importin
Exportin

ABSTRACT

Virus particle 24 (VP24) is the smallest protein of the Ebola and Marburg virus genomes. Recent experiments show that Ebola VP24 blocks binding of tyrosine-phosphorylated STAT-1 homodimer (PY-STAT1) to the NPI-1 subfamily of importin alpha, thereby preventing nuclear accumulation of this interferon-promoting transcription factor which, in turn, reduces the innate immune response of the host target. Lacking an experimental structure for VP24, we applied *de novo* protein structure prediction using the fragment assembly-based Rosetta method to classify its fold topology and better understand its biological function. Filtering and ranking of models were performed with the DFIRE all-atom statistical potential and the CHARMM22 force field with a generalized Born solvent model. From 40,000 Rosetta-generated structures and selective comparisons with the SCOP database, a structural match to two of our top 10-ranking models was the Armadillo repeat fold topology. Specific members of this fold family include importin alpha, importin beta, and exportin. We propose that, unlike the nuclear import of host cargo, VP24 lacks a classical nuclear localization signal (NLS) and targets importin alpha in a similar manner to the observed heterodimeric complex with exportin, thereby interfering with the auto-inhibitory NLS on importin alpha and blocking peripheral docking sites for PY-STAT1 assembly.

Published by Elsevier Inc.

1. Introduction

Ebola and Marburg hemorrhagic fever viruses, sole members of the *Filoviridae* family, produce two of the most deadly human diseases known. Of the seven proteins encoded in the viral genome, virus particle 24 protein (commonly referred to as "VP24") is the smallest and one of the least understood. As with many of the proteins in such a compactly-encoded virus, one might expect VP24 to exhibit more than one function. In fact, imaging studies of VP24-transfected cells (Han et al., 2003) indicate that the protein localizes both in the plasma membrane and near the nucleus. Roles in its membrane-bound state might include assistance with viral assembly (Huang et al., 2002), budding (Han et al., 2003), and transcription (Hoenen et al., 2006). In its soluble state, it serves as a matrix protein (Han et al., 2003). Moreover, experiments have shown that VP24 along with proteins VP35 and NP are necessary and sufficient to form nucleocapsid-like structures *in vivo* (Huang et al., 2002). Further studies concluded that VP24 is necessary for the nucleocapsid to be functional, as lack of VP24 in infectious virus-like particles led to reduced transcription and/or translation of another Ebola protein, VP30 (Hoenen et al., 2006). Somewhat paradoxically, VP24 also has been found to reduce transcription and replication of the virus (Watanabe et al., 2007).

Han et al. (2003) characterized different *in vitro* aqueous forms that may be related to its role as a matrix protein. Specifically, the molecular weight of VP24 oligomers was determined using different truncations of the N-terminus. The wild-type protein *in vitro* exists in both monomeric and tetrameric forms. Removing the first 40 residues disrupted tetramer formation and led to non-specific aggregation, suggesting that the N-terminus is critical for ordered self-polymerization.

A recent breakthrough was made in the elucidation of one of the functions of VP24 in the cytoplasm of the host. Reid et al. (2006) showed how Ebola VP24 acts in suppressing interferon production in host cells. VP24 binds to human importin α5, α6, and α7 (karyopherin α1, α5, and α6, respectively; i.e., the NPI-1 subfamily), preventing native tyrosine-phosphorylated STAT-1 homodimer (PY-STAT1) and STAT-1/STAT-2 heterodimer from translocating into the nucleus (Reid et al., 2006). PY-STAT1 and STAT-1/STAT-2 heterodimer are host-based transcriptional promoters of interferon α/β and γ, respectively.

In vitro studies suggest that VP24 competitively binds importin α5/α6/α7 displacing PY-STAT1 (Reid et al., 2006). If such a process were taking place, it is expected that VP24 binds to the same region of importin α5/α6/α7 as PY-STAT1, or it locks the auto-inhib-

* Corresponding author. Address: Department of Cell Biology and Biochemistry, US Army Medical Research Institute of Infectious Diseases, Frederick, MD 21702, USA. Fax: +1 301 619 2348.
E-mail address: michael.lee@amedd.army.mil (M.S. Lee).

1047-8477/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.jsb.2009.05.001

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
15 MAY 2009	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Fold prediction of VP24 protein of Ebola and Marburg viruses using de novo fragment assembly. J Struct Biol 167:136-144	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
Lee MS Lebeda FJ Olson MA	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
United States Army Medical Reseaarch Institute of Infectious Diseases, Fort Detrick, MD	**TR-08-009**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT
Virus particle 24 (VP24) is one of the l east understood proteins of the small genomes of Ebola and Marburg virus. Recent experiments show that Ebola VP24 blocks binding of tyrosine-phosphorylated STAT-1 homodimer (PY-STAT1) to the NPI-1 subfamily of importin alpha, thereby preventing nuclear accumulation of this interferon-promoting transcription factor which, in turn, reduces the immune response of the host target. Because of the lack of crystallographic structures for VP24, we applied a bioinformatics analysis and de novo protein structure prediction to classify its fold topology and better understand its biological function. Structure predictions were carried out using the fragment-based Rosetta method. Our scoring function for filtering and ranking structures consisted of a combined approach of an all-atom statistical potential and the CHARMM22 force field with a generalized Born solvent model. From 40,000 Rosetta-generated structures and selective comparisons with the SCOP database, a structural match to two of our top 10-ranking models was the Armadillo repeat fold topology, which is part of the alpha-alpha superhelix fold cluster. Specific members of the Armadillo repeat family include importin alpha, importinbeta, and exportin. We predict that a large segment of the VP24 topology is not unique, but rather represents an ancestral homolog of this existing repertoire of structures. We propose that, unlike the nuclear import of host cargo, VP24 lacks a classical nuclear localization signal (NLS) and targets importin alphaby structural complementarity similar to that observed with the heterodimeric complex with exportin, whereby interfering with the auto-inhibitory NLS on importin alphaand blocking peripheral docking sites for PY-STAT1 assembly. In addition to the dominant helical classification of VP24, a beta sheet is predicted near the C-terminus and may present an exposed coat protein segment responsible for the immunogenic properties reported for Ebola VP24.

15. SUBJECT TERMS
de novo prediction, Rosetta, fold recognition, filovirus, Ebola, Marburg, hemorrhagic fever virus, importin, exportin, VP24, computer modeling

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**SAR**	**9**	

itory NLS domain of importin α into the NLS-binding groove, preventing the NLS of PY-STAT1 from attaching. Melen et al. (2003) show from mutation experiments that the nuclear localization signal (NLS) of PY-STAT1 binds to the C-terminal region of importin α5. While it is not considered a traditional NLS-binding region, the structure of C-terminal-bound cargo was recently determined for an influenza PB2-importin α5 complex (PDB ID: 2JDQ) (Tarendeau et al., 2007).

Our study focused on predicting the structural topology of the VP24 monomer and its fold classification using bioinformatic and *de novo* approaches. Our predictions were evaluated in light of recent experimental observations of Ebola VP24 interfering with the nuclear import process. In addition, to further validate our *de novo* fold recognition protocol beyond studies performed previously (Bonneau et al., 2002), we performed fold prediction on another protein in the Ebola genome, VP30, whereby the structure of the C-terminal domain has been experimentally determined (Hartlieb et al., 2007).

2. Materials and methods

The conventional approach to structure prediction is to equate sequence stretches of a query protein with known protein structures using such tools as PSI-BLAST (Altschul et al., 1997). If this procedure fails, fold recognition/threading methods are used to deduce coarser-grained similarities of the estimated sequence characteristics with a database of known fold topologies. If both of these methods lead to inadequate confidence scores and/or a lack of consensus, computationally intensive *de novo* methods, such as fragment assembly (Simons et al., 1997), can be applied as a last resort. We performed all three types of procedures in this work.

First, we compared the sequences of Zaire Ebola VP24 (denoted here as Ebo-VP24; GenBank accession no. AAD14588; 251 aa) and Marburg VP24 (Mar-VP24; GenBank accession no. AAQ55260.1; 253 aa) against the non-redundant (NR) NCBI database (Benson et al., 2007) using PSI-BLAST (Altschul et al., 1997) and against the UNIPROT sequence database (Apweiler et al., 2004) using the Smith–Waterman algorithm (SSEARCH34) (Pearson, 1991; Smith and Waterman, 1981). For reference, the two VP24 sequences are pairwise aligned in Fig. 1. Next, we submitted both sequences to the Bioinfo.pl metaserver (Bujnicki et al., 2001) which invokes several fold recognition servers including Basic, MetaBasic (Ginalski et al., 2004), 3D-PSSM (Kelley et al., 2000), Orfeus (Ginalski et al., 2003b), FFAS03 (Jaroszewski et al., 2005), and the consensus algorithm, 3D-Jury (Ginalski et al., 2003a). In the third approach, we applied a *de novo* structure prediction strategy using the RosettaAbInitio program (version 1.1) (Simons et al., 1997), which has been successful in finding remote fold homologies when sequence-based methods fail (Bonneau et al., 2004). We used the program, DISpro v. 1.0 (Sickmeier et al., 2007), to identify potentially disordered regions in order to reduce the stretch of residues that were input to Rosetta.

Before assembling models with Rosetta, we constructed 3- and 9-residue fragment libraries using *make_fragments.pl* (Simons et al., 1997). Only one secondary structure prediction program, PSIPRED v. 2.4 (Jones, 1999), was used in this process. The server-based SAM-T06 secondary structure predictor (Karplus et al., 2005) produced very similar results. The fragment database "*vall-2001*" was used. PSI-BLAST (Altschul et al., 1997) invoked by the fragment library program discovered a weak sequence similarity (*e*-value = 0.4) to both Ebo-VP24 and Mar-VP24 in the fragment database: mouse importin α (PDB: 1IAL). As a result, we built two fragment libraries for each sequence, one with 1IAL and one without.

Twenty thousand backbone-only structures were generated by Rosetta for each sequence. Ten thousand of the models were generated using the 1IAL-included fragment library, another set of 10,000 structures were generated with the fragment library that excluded 1IAL. The SCWRL side-chain generation program (Canutescu et al., 2003) was used to generate all-atom models for each Rosetta-built backbone model. To improve conformational diversity (Bradley et al., 2005), another 20,000 models for each sequence were generated by substituting the side chains of the models built from one sequence with the side chains of the other sequence. In total, 40,000 all-atom models were generated for each VP24 sequence.

Each of the Rosetta models were evaluated with the DFIRE-based all-atom (DFIRE-AA) statistical potential using an in-house implementation (Zhang et al., 2004; Zhou and Zhou, 2002). While Rosetta has a built-in all-atom scoring function, DFIRE-AA is less sensitive to steric clashes and thus can be used on non-minimized models. The top-100 scoring structures for each batch of 10,000 models were minimized with 50 steps of steepest descent followed by 100 steps of adopted-basis Newton–Raphson using the *minCHARMM.pl* (Feig et al., 2004) which invokes CHARMM (Brooks et al., 1983). The energy function used for minimization was the PARAM22 force field (Mackerell et al., 1998) with a 4r-dielectric electrostatic function. Next, the PARAM22 plus generalized Born molecular volume solvation (GBMV2) (Lee et al., 2003) energy (including a 15 cal/mol/Å^2 nonpolar surface area term) was evaluated for each batch of 100 minimized structures.

Early papers describing Rosetta methodology employed clustering of generated models to extract out representative models (Tsai et al., 2003). Our studies (Lee and Olson, 2007) and more recent

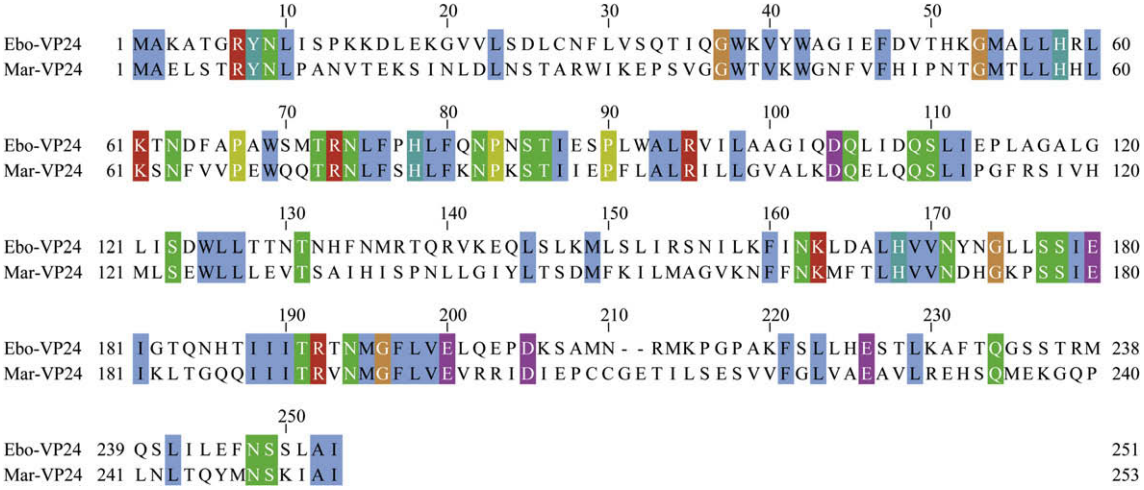

Fig. 1. Sequence alignment of Ebo-VP24 and Mar-VP24 rendered with Jalview 2.4 (Waterhouse et al., 2009). The highlighted residues are conserved between the two species.

Table 1
Notable high-scoring models from the Bioinfo.pl metaserver for VP24 (Bujnicki et al., 2001).

Annotation	PDB ID (chain)	SCOP code	Method	Scoring method	Score	VP24 strain[a]	Sequence range[b]
Chaperonin 60	1iok(A)	a.129.1.1	Basic(dist) (Ginalski et al., 2004)	3D-Jury (Ginalski et al., 2003a)	39.00/ 37.67	E/M	23–200
DNA repair protein	1ee8(A)	a.156.1.2	Basic(dist)	3D-Jury	25.14/28.5	E/M	66–163
Cytochrome b6f	1q90(D)	f.32.1.1[c]	FFAS03 (Jaroszewski et al., 2005)	3D-Jury	14.43/16.5	E/M	54–191
Nuclear cap binding protein	1n52(A)	a.118.1.14	Basic(dist)	3D-Jury	14.29	E	53–162
IRF-3	1j2f(A)	b.26.1.3[d]	FFAS03	FFAS03	−6.8/−6.6	E/M	15–164

[a] E = Ebola VP24 sequence; M = Marburg VP24.
[b] VP24 sequence range that was aligned to target sequence.
[c] Transmembrane helix topology.
[d] All-beta topology.

work in the Baker lab (Bradley et al., 2005; Misura and Baker, 2005) suggest that it is better to attempt to detect the rare models that achieve the lowest scores rather than clustering. The rationale is that the united-residue energy function used in the *de novo* fragment assembly phase is lower resolution and may not significantly populate near-native topologies. Instead, if a rare near-native model is generated, the expectation is that an all-atom scoring function will select it out (Bradley et al., 2005; Lee and Olson, 2007).

Finally, the five top scoring structures from all of the batches for each sequence were queried in a structural similarity search against the 95% homologous ASTRAL subset (Chandonia et al., 2004) of the SCOP 1.71-fold database (Andreeva et al., 2004; Murzin et al., 1995) using the Combinatorial Extension (CE) program (Shindyalov and Bourne, 1998). The CE *Z-score* is an aggregate measure of the RMSD, length, and gaps of the optimal alignment between query and template. *Z*-score values above 4.5 are indicative of fold similarity; values above 5 are more compelling. The CE program was slightly modified for this work to generate continuous *Z*-scores from the continuous probability, *P*:

$$Z\text{-score} = 1.26098\sqrt{-3\log_{10}P} - 0.445628,$$

derived by a fit to the tabular data in the source code.

We have already benchmarked our post-processing scoring functions on Rosetta-generated models with less than 100 residues (Lee and Olson, 2007). Furthermore, others have shown that it is possible to deduce fold type from *de novo* predictions (Bonneau et al., 2004; Bonneau et al., 2002). Nonetheless, we include in this work a validation of our *de novo* fold recognition protocol on the C-terminal domain (residues 140–266) of Ebola VP30 which has already been structurally characterized (PDB ID: 2I8B) (Hartlieb et al., 2007). The C-terminus of VP30 has a unique fold amongst known folds and, like VP24, is dissimilar to any other known sequence outside of filoviruses. The specific sequences evaluated in this work were GenBank accession no. AAD14587 (“Ebo-VP30”; residues 140–266) and GenBank accession no. AAQ55259 (“Mar-VP30”; residues 147–274). There were two differences in our *de novo* prediction of this protein vs. VP24. First, we did not substitute the side chains of the Ebola sequence in the Marburg models and vice versa. Thus, only 20,000 models per sequence were generated. Second, because there were no homologous fragments detected, the 10,000-model runs were simply performed in duplicate.

3. Results

Prior to obtaining fold recognition and *de novo* predictions, several analyses on the primary structures and sequences were conducted. We initially considered possible unstructured regions in VP24. The first 20 residues of the N-terminal and the last 50 C-terminal residues for both Ebola and Marburg VP24 sequences were predicted by the DISpro program (Sickmeier et al., 2007) to have at least some structural disorder. The N-terminus of VP24 is known to induce tetramers as opposed to high molecular-weight aggregates (Han et al., 2003). Conceivably, the N-terminus region becomes more structured as part of the tetrameric binding interface.

We next examined secondary structure predictions for the entire Ebo-VP24 and Mar-VP24 sequences using PSIPRED version 2.45 (Jones, 1999). The proteins are nearly equivalent in this analysis as the position-specific substitution matrices are each composed of the same set of filoviral proteins. Residues 50–175 appear to be an all-α arrangement, while residues 175–203 form a contiguous multi-stranded β-sheet.

Optimal pairwise sequence alignments (SSEARCH) with the UNIPROT sequence database (Apweiler et al., 2004) and PSI-BLAST alignments (Altschul et al., 1997) with the NR protein database (Benson et al., 2007) indicate that both Ebo-VP24 and Mar-VP24 have no clear sequence homologues to any other known viral, prokaryotic, and eukaryotic sequences available so far. In contrast, the VP24 sequences from the two species are 35% identical and represent the most conserved viral proteins within the *Filoviridae* family.

3.1. Fold recognition servers

Similar to the sequence searches, fold recognition programs instantiated from the Bioinfo (http://bioinfo.pl) metaserver for VP24 provided relatively low confidence model predictions. In Table 1, the most notable and representative structural neighbors predicted include chaperonins, DNA repair proteins, cytochromes, nuclear cap binding protein and IRF-3. From the SCOP database, the corresponding detected fold types were the following: multihelical, consisting of 8 helices arranged in two parallel layers; 3–4 helices; membrane and cell-surface proteins with three transmembrane helices forming an up-and-down bundle; α–α superhelix; and all-β sandwich. As discussed further below, the prevalence of protein classification was an all-α bundle with non-orthogonal helices.

3.2. De novo predictions

Lacking clear consensus from the sequence-based fold recognition servers, we applied a *de novo* structure approach to find remote fold homologs (Bonneau et al., 2002, 2004). Since the Rosetta folding program is indicated for protein segments of less than 200 residues (Chivian et al., 2003), we limited our study to a 151 residue stretch. Using the most conservative measure from DISpro, residues 21–201 had no predicted disorder. In addition, as seen in Fig. 1, the sequence region 50–200, which we chose to model via *de novo* fragment assembly, is moderately conserved

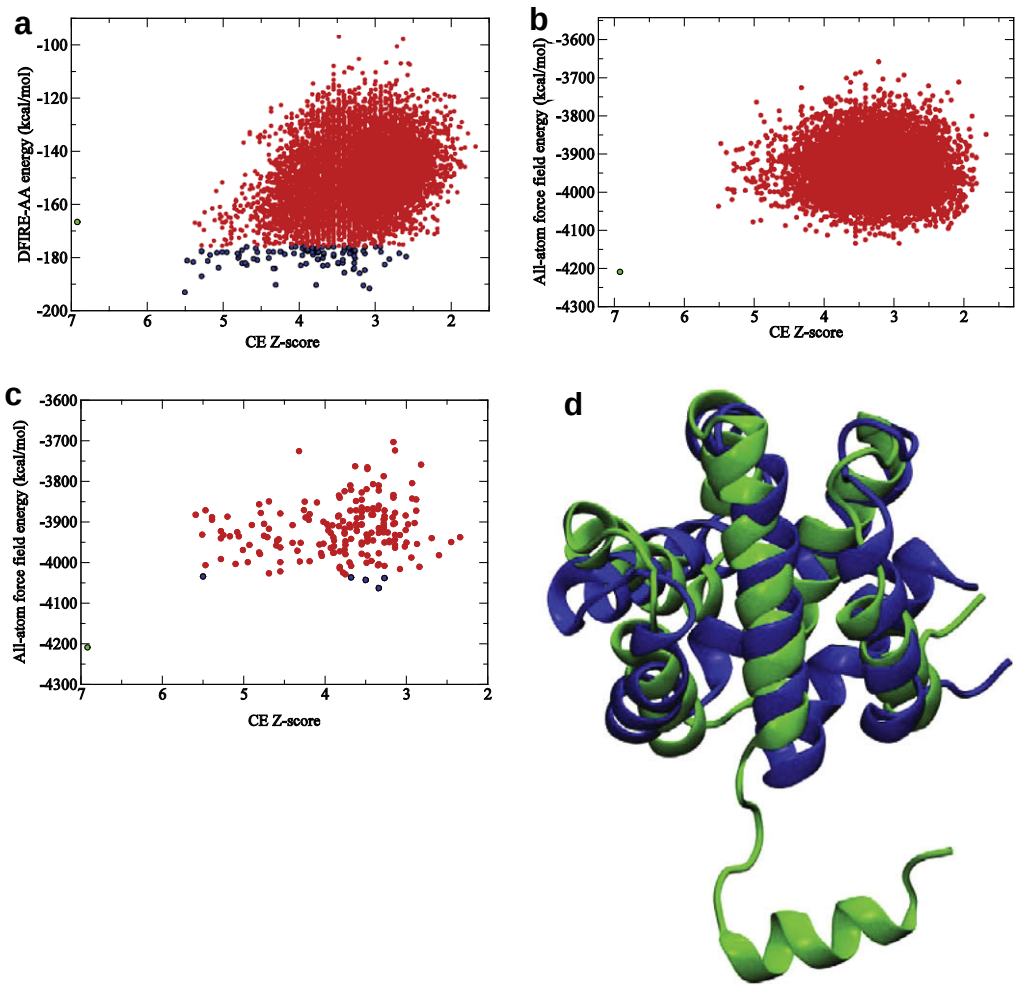

Fig. 2. *De novo* fragment assembly of Ebo-VP30. (a) DFIRE-AA energies vs. pairwise CE *Z*-scores to native VP30 C-terminal domain (PDB ID: 2I8B) (Hartlieb et al., 2007) for all Rosetta models generated in the 2nd run. (b) All-atom force field energies vs. CE *Z*-score for all Rosetta models generated in the 2nd run. (c) All-atom force field energies vs. CE Z-score of the 200 models collected from the top-100 DFIRE-AA structures from each run. (d) Graphical representation of the most native-like Ebo-VP30 model (blue) from the list of the top-5 scoring models superimposed on the experimental structure (green) (C_α RMSD = 4.5 Å for residues 140–250).

(45% similarity) between the sequences of Ebo-VP24 and Mar-VP24 with no alignment gaps. As a confirmation of our region of focus, *de novo* predictions based on regions 1–150, 100–251 (Ebo-VP24), and 100–253 (Mar-VP24) yielded little consensus among ascertained fold types (results not shown).

To select protein models from Rosetta for structural comparisons, we applied a multi-resolution scoring and filtering approach. Since our model scoring approach deviates from the Rosetta all-atom refinement protocol (Bradley et al., 2005; Lee and Olson, 2007), we first illustrate the procedure with the sequence stretch of Ebo-VP30 which has a known structure in Fig. 2 and Table 2. In Fig. 1a, the DFIRE-AA energy has a moderate scoring funnel compared to the CE structural similarity measure. On the other hand, Fig. 2b shows that the all-atom CHARMM22/GBMV2 energy does not have a scoring funnel with the models generated, albeit the native structure is the lowest scoring by a significant margin. Using DFIRE-AA as a filter, by retaining only the two top-100 DFIRE-AA subsets, the CHARMM22/GBMV2 energy selects out one model in the top-5 with a CE *Z*-score of 5.4 compared to the native structure. This model is superimposed on the experimental structure in Fig. 2d. The C_α RMSD between the two structures is 4.5 Å in the region 140–250. The C-terminal segment (251–270) is not compacted to the rest of the structure because it is bound to an identical chain in the experimentally-determined dimer (Hartlieb et al., 2007). Finally, Table 2 summarizes the structural similarity searches of the top-ranking Rosetta models against the SCOP database of known folds. The search consensus between Ebo-VP30 and Mar-VP30 correctly selects out the VP30 fold. While not an entry in SCOP 1.71, the VP30 fold is unique (Hartlieb et al., 2007).

Table 2
Top SCOP fold type matches (as measured by CE structural similarity *Z*-score) for the top five scoring Ebo-VP30 and Mar-VP30 Rosetta protein structure models.

Fold type/family	SCOP code	Ebola models	Marburg models	Average
VP30	n/a	5.43	5.12	5.28
Delta-endotoxin, N-terminal	f.1.3.1	5.12	4.93	5.03
Four-helical up-and-down bundle/α-catenin(vinculin)	a.24.9.1	5.06	4.98	5.02
Nickel-containing superoxide dismutase	a.24.22.1	4.46	4.98	4.72

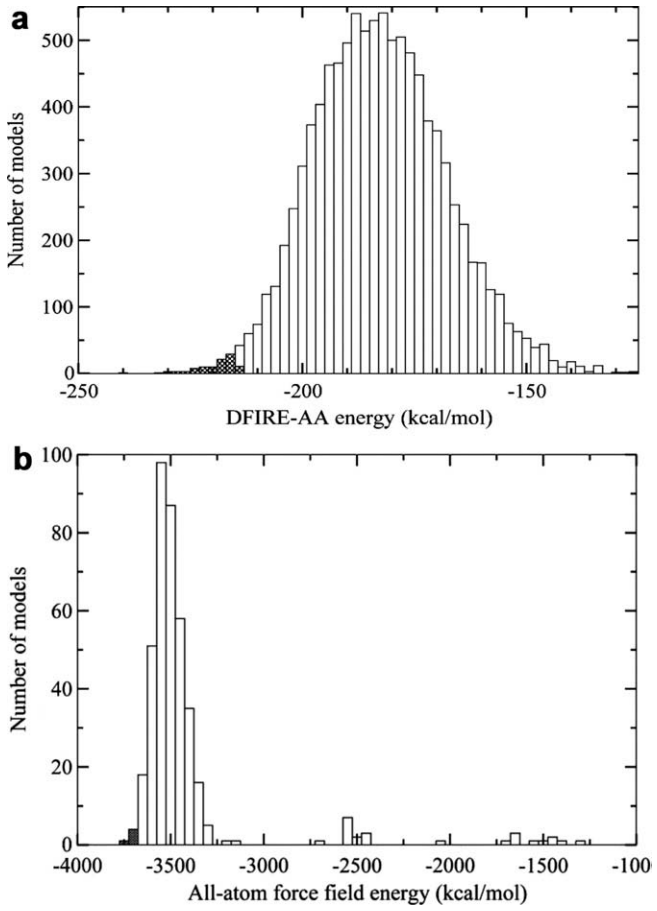

Fig. 3. (a) Distribution of DFIRE-AA energies for a single batch of 10,000 all-atom Rosetta models of Ebo-VP24 (residues 50–200). The shaded bars indicate the 100 top-scoring models selected for further minimization and scoring with CHARMM/GB. (b) Distribution of CHARMM/GB scores for the conglomerated 400 minimized models of Ebo-VP24 (residues 50–200). The shaded bars, indicate the five top-scoring models which were submitted to the CE structural similarity search.

Proceeding to the blind prediction of VP24, shown in Fig. 3 are the energy distributions for the structures generated for Ebo-VP24. Fig. 3a illustrates DFIRE-AA energies for the 10,000 Rosetta models of Ebo-VP24 from which 100 high-ranking structures (shaded bars) were selected for further scoring. A total of 40,000 all-atom models for each sequence were generated and a total of 400 structures were selected from each of four 10,000-model sets (see Methods). Fig. 3b illustrates the application of the CHARMM22/GBMV2 method to select the top five models (shaded bars) out of the 400 used in the CE structure–structure alignment similarity search.

Listed in Table 3 are the top scoring hits for the VP24 proteins with *Z*-scores > 4.5 from the CE analysis. Because of the 35% sequence identity between Ebo-VP24 and Mar-VP24, we once again have taken a consensus approach of ranking the structural neighbors of the five selected Rosetta models. The results show the top hit was the Armadillo repeat fold family, which includes proteins importin α, importin β, exportin and β-catenin. As noted above, importin α has been shown to be a host target for Ebo-VP24.

Illustrated in Fig. 4 are three of the top-ranking structural models from Rosetta and their 3-D protein topology diagrams (Westhead et al., 1998). Regions of the Rosetta models that structurally align with known protein structures are colored green and regions lacking alignments are colored blue for the N-terminal and red for the β-strand regions. For each of the Rosetta models, the depicted topology cartoons reflect the diversity among the three predicted folds in packing helices (colored green) and β-strands (red).

4. Discussion

Our prediction protocols were not able to obtain a consensus structural model of the Ebo-VP24 and Mar-VP24 proteins based entirely on sequence homology or fold recognition. Of the latter, four of the five top-scoring SCOP classifications were all-α proteins, including a transmembrane helical up-and-down bundle structure which is consistent with the viral matrix nature of VP24. The remaining predicted model was an all-β fold. Although there was strong consensus of predicting all-α proteins, many different folds were recognized of this classification. To help place this prediction in its proper perspective, there are nearly 1100 unique protein folds classified in SCOP and of this total nearly 260 are all-α helical proteins (Andreeva et al., 2004).

In an attempt to improve the structural classification of the Ebo-VP24 and Mar-VP24 proteins beyond sequence and fold threading, we applied *de novo* predictions using the Rosetta fragment assembly method. A multi-resolution scoring approach was then applied to filter and select protein models for structural comparisons with known folds of the SCOP library. The application of the DFIRE-AA potential to the Rosetta-generated structures yielded a distribution of energies that presumably favored a small number of native or near-native structures from decoys. The top-ranking DFIRE-AA structures were then minimized slightly to relieve steric clashes. More extensive refinement, using for example the Rosetta model relaxation module, can be very constructive when at least some of the models are expected to be very close to the native structure (Bradley et al., 2005). However, extensive refinement is significantly more costly than simple minimization, and, perhaps, needlessly expensive for the lower-resolution fold prediction sought in this work. The CHARMM22/GBMV scoring function differentiates among the top-ranking DFIRE structures and provides detection of protein models that exhibit higher-resolution local geometric properties (e.g., all-atom side-chain packing, electrostatic solvent effects, etc.). The optimal goal of any scoring function is to assess sequence fitness for a given protein fold and our combined approach narrowed the selection down to five protein models from a total of 40,000 structures for each sequence.

Other researchers have shown that top-scoring Rosetta-generated models can indeed be used to identify fold type through structural similarity (Bonneau et al., 2002, 2004). However, since our post-processing algorithm is technically different, we first presented results for another Ebola protein, VP30, to validate our *de*

Table 3
Top SCOP fold type matches (as measured by CE structural similarity *Z*-score) for the top five scoring Ebo-VP24 and Mar-VP24 Rosetta protein structure models.

Fold type/family	SCOP code	Ebola models	Marburg models	Average
α–α Superhelix/Armadillo repeat	a.118.1.1	4.87	5.43	5.15
α–α Superhelix/MIF4G domain-like	a.118.1.14	4.53	5.45	4.99
Four-helical up-and-down bundle/mannose-6-phosphate receptor binding protein	a.24.23.1	5.27	4.67	4.97
Four-helical up-and-down bundle/α-catenin(vinculin)	a.24.9.1	4.80	4.80	4.80
Cytochrome P450/cytochrome P450	a.104.1.1	4.65	4.85	4.75

novo protocol with a known protein structure in the same genome with a similar number of residues ($N = 127$) (Bradley et al., 2005; Lee and Olson, 2007). What we see from this experiment is that DFIRE-AA acts as a low resolution filter to cull out models moderately similar to the native. While it has been shown elsewhere that CHARMM22/GBMV2 energy can perform a higher resolution detection of near-native models than DFIRE-AA (Lee and Olson, 2007), it appears in this case that CHARMM22/GBMV2 did not enhance detection, presumably because the Rosetta models were not close enough to the native. Encouragingly, we were able to correctly identity of the fold type of VP30 (Table 2). However, had the experimentally-determined VP30 fold not been present, we would have had difficulty classifying the sequence as having a new fold.

Comparisons of our selected models from Rosetta for VP24 with structures from SCOP revealed several interesting results. Using a consensus approach of treating the Ebola and Marburg proteins as evolutionarily conserved topologies, the top-scoring models for residues 50–200 were observed to populate two major fold clusters: α–α superhelix (rank 1 and 2) and four-helical up-and-down bundle (rank 3 and 4). There are multiple fold topologies for the α–α superhelix lineage and the SCOP database contains 23 fold superfamilies for this cluster. For the four-helical up-and-down bundle, there are 27 superfamilies. While the average

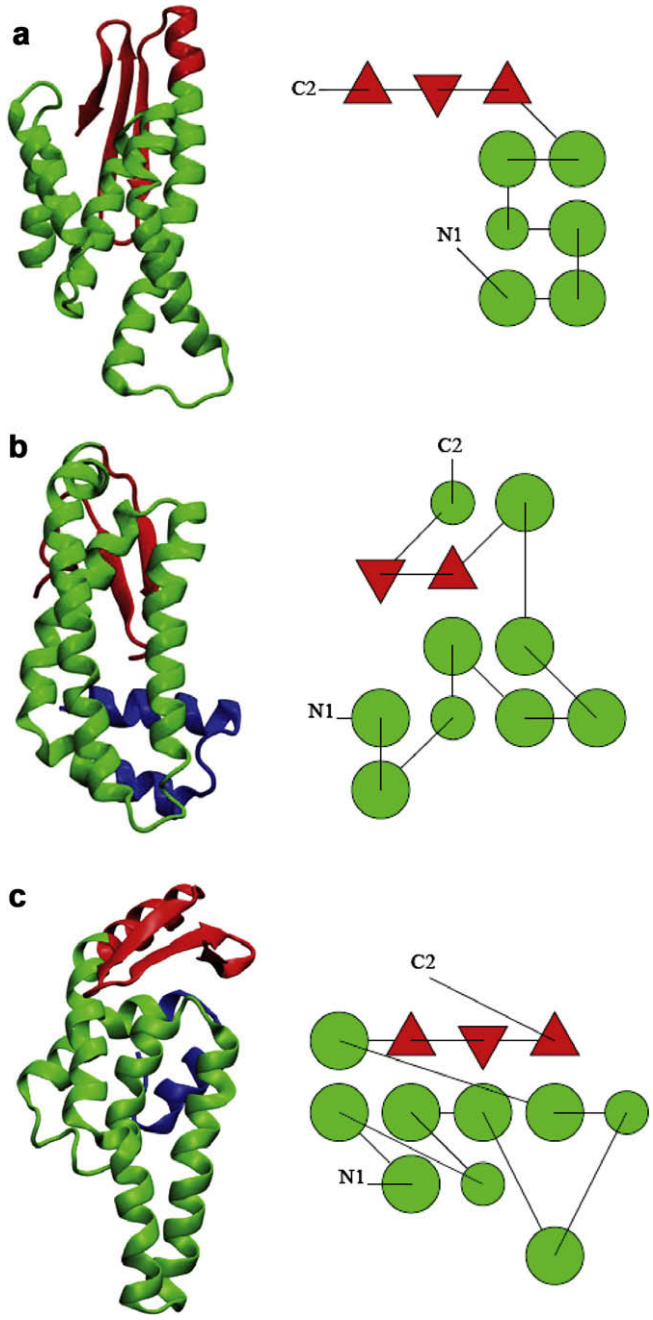

Fig. 4. Three of the top-scoring VP24 structural models (residues 50–200) and their corresponding TOPS (Westhead et al., 1998) topology diagrams on the right: (a) Armadillo repeat fold (α-helical segments: 52–59, 68–79, 85–101, 104–130, 137–172; β-sheet segments: 177–183, 187–193, 196–199); (b) superantigen MAM fold (α-helical segments: 52–62, 69–81, 84–87, 89–98, 104–110, 112–125, 135–158, 160–173; β-sheet segments: 177–180, 187–190; 3_{10}-helical segment: 195–197); and (c) alpha-catenin/vinculin fold (α-helical segments: 55–62, 66–73, 76–81, 85–88, 94–100, 105–127, 131–141,147–160,161–173; β-sheet segments: 177–182, 187–192, 197–199). Segments labeled by helical content (green), sheet content (red), and region not aligned with structural homologue (blue).

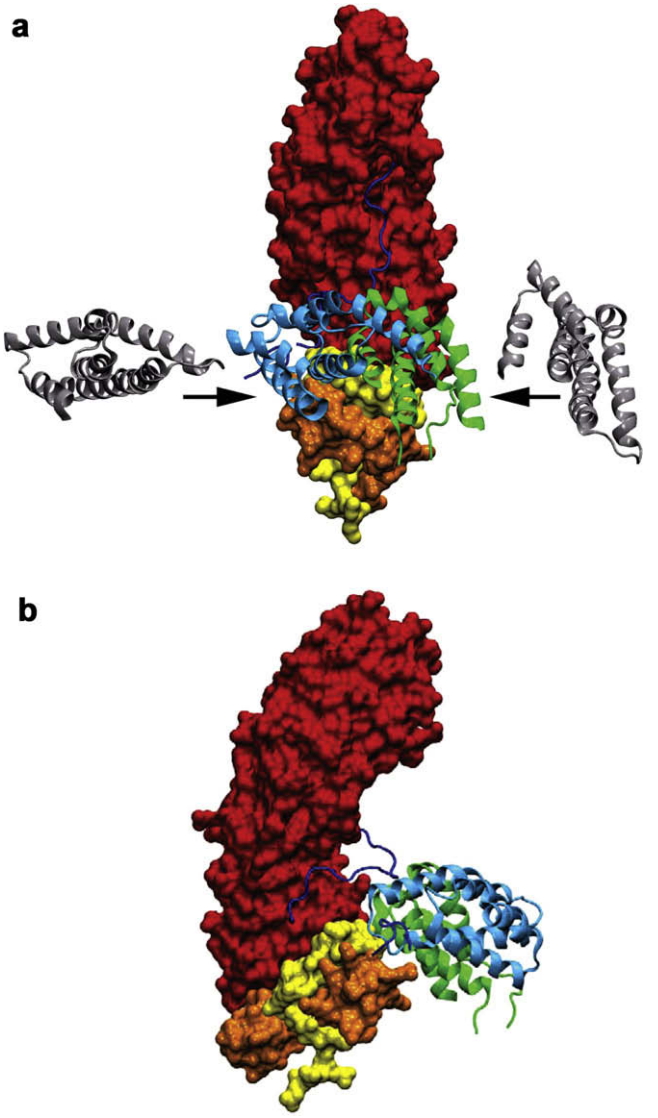

Fig. 5. X-ray crystal structure (PDB ID: 1WA5) (Matsuura and Stewart, 2004) of importin α (red, orange, and yellow molecular surfaces (Sanner et al., 1996)) and its auto-NLS (dark blue tube) bound to fragments of exportin (light blue ribbons [residues 84–190] and green ribbons [residues 211–319]): (a) front view; (b) side view. The Armadillo-fold VP24 model (silver ribbons) in (a) structurally overlaps the first exportin fragment with high confidence (Z-score = 5.2) and overlaps the second exportin fragment with moderate confidence (Z-score = 4.2). Orange-colored region of importin α was experimentally identified to be minimally necessary for VP24 binding (Reid et al., 2007). The combined yellow and orange regions indicate the experimentally characterized PY-STAT1 binding region (Melen et al., 2003; Reid et al., 2007).

Fig. 6. Sequence and secondary structure comparison of residues 51–170 of Armadillo-fold VP24 model with residues 85–192 of exportin Cse1p (PDB ID: 1WA5) using the structural alignment obtained by CE (*Z*-score = 5.2, RMSD = 5.2 Å, sequence identity = 7.4%). The secondary structure assignments were generated with DSSP (Kabsch and Sander, 1983). DSSP legend: H – alpha helix, G – 3_{10} helix, T – hydrogen-bonded turn, S – bend, and C – loop (no ordered structure).

Z-scores separating the structural models are less than 0.5 (Table 2), individual *Z*-scores of approximately 5 or greater are very robust in detecting structural neighbors. Moreover, the set of scores computed among the structural alignments for each protein model showed clear discrimination of rank 1 thru 4 within a fold space of 50 helical superfamilies.

The rank 1 of Armadillo repeat family is the most intriguing result for the classification of VP24. What is convincing about this hit is that this fold family contains importin α and β, plus exportin. As reported from experimental studies, Ebo-VP24 targets importin α and inhibits intracellular signaling of the interferon pathway. Specifically, VP24 interferes with the association of PY-STAT1 to the C-terminus of importin α5/α6/α7. We predict that there is no fold similarity between VP24 and PY-STAT1, but rather that the viral protein is a distant structural homolog of molecules that participate in transporting cargo across the nuclear envelope.

Essential to deriving a model of VP24 are experimental observations of how importin α recognizes and binds other proteins, including members of the same superfamily. The most common way that proteins bind importin α is through the use of nuclear localization signals (NLS), which are short stretches of residues with multiple basic amino acids (i.e., lysine and arginine). Indeed, Ebo-VP24 contains a curious segment of residues, 13-PKKDLEK-19, which is almost completely conserved among its sequence variants (results not shown). However, this region does not produce any matches with the NLSdb (Nair et al., 2003), which is a database of known NLS motifs. Furthermore, this stretch of residues is completely absent in Mar-VP24. If this stretch of residues in Ebo-VP24 were binding like an NLS, and this was its sole mode of binding to importin α, then it would not be possible to explain how VP24 prevents PY-STAT1 and STAT-1/STAT-2 heterodimer from translocating into the nucleus, given that NLS sequences, which are quite common in the proteome of a cell, are not known to inhibit the nuclear translocation process. Therefore, we can eliminate the possibility that VP24 uses only an NLS to form the association complex.

Alternatively, VP24 must bind importin α independent of NLS binding. Experimental models of protein–protein complex formation of importin α that are independent of NLS binding are homodimerization (e.g., PDB ID 2JDQ) (Tarendeau et al., 2007) and the association with exportin in the yeast proteome (PDB ID: 1WA5) (Matsuura and Stewart, 2004). Although homodimerization of importin α is observed under crystallographic conditions, no experimental data exist showing any biological significance and thus this model of binding VP24 can be presumably ruled out. In contrast, the binding of exportin recycles importin α from the nucleus back to the cytoplasm. Based on the crystallographic structure of importin α bound with exportin and using structure–structure alignments of our highest-ranked VP24 structure with exportin, we propose a model illustrated in Fig. 5. The docking sites for VP24 are located where the structural alignments with the ARM repeats of exportin showed the greatest interfacial contacts with importin α. Our model predicts that one or more monomers of VP24 bind importin α with structural complementarity to that observed with exportin (Fig. 5). In fact, the binding interface of exportin–importin complex overlaps with ARM repeat 10 (residues 458–504), which is the region on importin α experimentally known to be required for VP24 binding (Reid et al., 2007).

There are two consequences of this VP24-importin α structural model. First, the docked VP24 structures may hinder the release of the auto-inhibitory NLS of importin α, which can prevent NLS's from other proteins such as PY-STAT1 from docking into the NLS-binding groove. Subsequently, the N-terminus of importin α locked in its inhibitory state would be unable to bind importin β (Cingolani et al., 1999) and subsequently transport cargo such as PY-STAT1. Second, placement of VP24 at the predicted docking sites would preclude PY-STAT1 from binding to importin α in the vicinity of residues 425–538, which is the section of importin α known to be minimally required for PY-STAT1 complexation (Reid et al., 2007). For either of these scenarios to occur, the binding affinity of monomeric or multimeric VP24 to importin α must be fairly strong. For reference, the binding affinity of exportin to importin is ~1 nm (Kutay et al., 1997), where RanGTP is a necessary component for complexation. It is not possible to deduce specific pairwise residue contacts between VP24 and importin because the CE structural alignments of our VP24 model with the helical repeats of the exportin structure (Fig. 5) yield a sequence identity of 7.4% as seen in Fig. 6. Nonetheless, the secondary structural elements of the two structures align fairly well, further validating the fold similarity between our VP24 model and exportin. Further complicating detailed binding pattern assignments is the fact that the exportin/importin structure is the yeast variant and not the filovirus-relevant human version.

While our prediction suggesting that VP24 has ancestral links to the Armadillo repeat family is appealing, our model also shows divergence from this superfamily by containing a β-sheet arrangement packed against the helical bundle. The prediction accuracy of a β-sheet is strong among the top-scoring models (Fig. 4) and may highlight the multi-functional nature of VP24. Because of the lack of this secondary structure in exportin, the question becomes is there any functional role of the β-sheet in VP24? One possibility is stabilization of the VP24 monomer in aqueous solution, where apolar residues of the β-sheet shield the hydrophobic core from the aqueous environment while the hydrophilic residues are positioned toward the solvent. Many of these residues are conserved between Ebola and Marburg. Among the top-scoring models, the β-sheet is connected to the helical core by a common "hinge" segment of a polar residue and a glycine. When submerged in the viral lipid matrix, the β-sheet may disconnect from the hydrophobic face of the now-transmembrane helices (predicted in Table 1) as exposure of this surface becomes favored. In addition, this hinge may be active upon binding exportin (Fig. 5) as well as the self-ordered polymerization of VP24 in aqueous solvent.

Overall our model of the VP24 protein provides testable hypotheses for additional experimental work. Using mutagenesis, the docking sites on importin α that are predicted to overlap between VP24 and exportin can be examined to validate our binding model. Mutation studies should also include deleting the auto-inhibitory NLS segment of importin α to determine whether this affects binding of VP24. In a similar fashion, the β-sheet region suggests a new target to explore protein stability. Finally, our VP24 structural model and the fold similarity with exportin may prove to be helpful in the design of modifications that overcome the obstacles to the crystallographic determination of this viral protein critical to understanding Ebola virus infections.

Acknowledgments

We thank Drs. S. Bavari, R. Harty, A. Wallqvist, and K. Erickson for helpful discussions. Financial support for this work comes from DoD Defense Threat Reduction Agency grant (DTRA 4.10011_07_RD_B to MAO), from the Computational Biology program of the DoD Advanced Research Project Agency (DARPA 05-0-DA-008 to FJL) and from the Department of Defense High Performance Computing Modernization Program Office (HPCMO) and the Biotechnology High Performance Computing Software Applications Institute (BHSAI) to MSL. We thank the Army Research Laboratory DoD Supercomputing Resource Center for computer time. The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the US Army or of the US Department of Defense. This paper has been approved for public release with unlimited distribution.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C., Murzin, A.G., 2004. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res. 32, D226–229.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., Yeh, L.S., 2004. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 32, D115–119.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L., 2007. GenBank. Nucleic Acids Res. 35, D21–25.

Bonneau, R., Baliga, N.S., Deutsch, E.W., Shannon, P., Hood, L., 2004. Comprehensive de novo structure prediction in a systems-biology context for the archaea Halobacterium sp. NRC-1. Genome Biol. 5, R52.

Bonneau, R., Strauss, C.E., Rohl, C.A., Chivian, D., Bradley, P., Malmstrom, L., Robertson, T., Baker, D., 2002. De novo prediction of three-dimensional structures for major protein families. J. Mol. Biol. 322, 65–78.

Bradley, P., Misura, K.M., Baker, D., 2005. Toward high-resolution de novo structure prediction for small proteins. Science 309, 1868–1871.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminatham, S., Karplus, M., 1983. CHARMm: a program for macromolecular energy, minimization, and dynamics calculations. J. Comp. Chem. 4, 187.

Bujnicki, J.M., Elofsson, A., Fischer, D., Rychlewski, L., 2001. Structure prediction meta server. Bioinformatics 17, 750–751.

Canutescu, A.A., Shelenkov, A.A., Dunbrack Jr., R.L., 2003. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci. 12, 2001–2014.

Chandonia, J.M., Hon, G., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M., Brenner, S.E., 2004. The ASTRAL compendium in 2004. Nucleic Acids Res. 32, D189–192.

Chivian, D., Kim, D.E., Malmstrom, L., Bradley, P., Robertson, T., Murphy, P., Strauss, C.E., Bonneau, R., Rohl, C.A., Baker, D., 2003. Automated prediction of CASP-5 structures using the Robetta server. Proteins 53 (Suppl. 6), 524–533.

Cingolani, G., Petosa, C., Weis, K., Muller, C.W., 1999. Structure of importin-beta bound to the IBB domain of importin-alpha. Nature 399, 221–229.

Feig, M., Karanicolas, J., Brooks 3rd, C.L., 2004. MMTSB tool set: enhanced sampling and multiscale modeling methods for applications in structural biology. J. Mol. Graph. Model. 22, 377–395.

Ginalski, K., Elofsson, A., Fischer, D., Rychlewski, L., 2003a. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19, 1015–1018.

Ginalski, K., von Grotthuss, M., Grishin, N.V., Rychlewski, L., 2004. Detecting distant homology with Meta-BASIC. Nucleic Acids Res. 32, W576–581.

Ginalski, K., Pas, J., Wyrwicz, L.S., von Grotthuss, M., Bujnicki, J.M., Rychlewski, L., 2003b. ORFeus: detection of distant homology using sequence profiles and predicted secondary structure. Nucleic Acids Res. 31, 3804–3807.

Han, Z., Boshra, H., Sunyer, J.O., Zwiers, S.H., Paragas, J., Harty, R.N., 2003. Biochemical and functional characterization of the Ebola virus VP24 protein: implications for a role in virus assembly and budding. J. Virol. 77, 1793–1800.

Hartlieb, B., Muziol, T., Weissenhorn, W., Becker, S., 2007. Crystal structure of the C-terminal domain of Ebola virus VP30 reveals a role in transcription and nucleocapsid association. Proc. Natl. Acad. Sci. USA 104, 624–629.

Hoenen, T., Groseth, A., Kolesnikova, L., Theriault, S., Ebihara, H., Hartlieb, B., Bamberg, S., Feldmann, H., Stroher, U., Becker, S., 2006. Infection of naive target cells with virus-like particles: implications for the function of Ebola virus VP24. J. Virol. 80, 7260–7264.

Huang, Y., Xu, L., Sun, Y., Nabel, G.J., 2002. The assembly of Ebola virus nucleocapsid requires virion-associated proteins 35 and 24 and posttranslational modification of nucleoprotein. Mol. Cell. 10, 307–316.

Jaroszewski, L., Rychlewski, L., Li, Z., Li, W., Godzik, A., 2005. FFAS03: a server for profile–profile sequence alignments. Nucleic Acids Res. 33, W284–288.

Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195–202.

Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637.

Karplus, K., Katzman, S., Shackleford, G., Koeva, M., Draper, J., Barnes, B., Soriano, M., Hughey, R., 2005. SAM-T04: what is new in protein-structure prediction for CASP6. Proteins 61 (Suppl. 7), 135–142.

Kelley, L.A., MacCallum, R.M., Sternberg, M.J., 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J. Mol. Biol. 299, 499–520.

Kutay, U., Bischoff, F.R., Kostka, S., Kraft, R., Gorlich, D., 1997. Export of importin alpha from the nucleus is mediated by a specific nuclear transport factor. Cell 90, 1061–1071.

Lee, M.S., Olson, M.A., 2007. Assessment of detection and refinement strategies for de novo protein structures using force field and statistical potentials. J. Chem. Theory Comput. 3, 312–324.

Lee, M.S., Feig, M., Salsbury Jr., F.R., Brooks 3rd, C.L., 2003. New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. J. Comput. Chem. 24, 1348–1356.

Mackerell Jr., A.D., Bashford, D., Bellott, D.M., Dunbrack Jr., R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F.T.K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D.T., Prodhom, B., Reiher, I., Roux, W.E.B., Schlenkrich, M., Smith, J.C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D., Karplus, M., 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B 102, 3586–3616.

Matsuura, Y., Stewart, M., 2004. Structural basis for the assembly of a nuclear export complex. Nature 432, 872–877.

Melen, K., Fagerlund, R., Franke, J., Kohler, M., Kinnunen, L., Julkunen, I., 2003. Importin alpha nuclear localization signal binding sites for STAT1, STAT2, and influenza A virus nucleoprotein. J. Biol. Chem. 278, 28193–28200.

Misura, K.M., Baker, D., 2005. Progress and challenges in high-resolution refinement of protein structure models. Proteins 59, 15–29.

Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540.

Nair, R., Carter, P., Rost, B., 2003. NLSdb: database of nuclear localization signals. Nucleic Acids Res. 31, 397–399.

Pearson, W.R., 1991. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith–Waterman and FASTA algorithms. Genomics 11, 635–650.

Reid, S.P., Valmas, C., Martinez, O., Sanchez, F.M., Basler, C.F., 2007. Ebola virus VP24 proteins inhibit interaction of NPI-1 subfamily karyopherin {alpha} proteins with activated STAT1. J. Virol..

Reid, S.P., Leung, L.W., Hartman, A.L., Martinez, O., Shaw, M.L., Carbonnelle, C., Volchkov, V.E., Nichol, S.T., Basler, C.F., 2006. Ebola virus VP24 binds karyopherin alpha1 and blocks STAT1 nuclear accumulation. J. Virol. 80, 5156–5167.

Sanner, M.F., Olson, A.J., Spehner, J.C., 1996. Reduced surface: an efficient way to compute molecular surfaces. Biopolymers 38, 305–320.

Shindyalov, I.N., Bourne, P.E., 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. 11, 739–747.

Sickmeier, M., Hamilton, J.A., LeGall, T., Vacic, V., Cortese, M.S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V.N., Obradovic, Z., Dunker, A.K., 2007. DisProt: the database of disordered proteins. Nucleic Acids Res. 35, D786–793.

Simons, K.T., Kooperberg, C., Huang, E., Baker, D., 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J. Mol. Biol. 268, 209–225.

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197.

Tarendeau, F., Boudet, J., Guilligay, D., Mas, P.J., Bougault, C.M., Boulo, S., Baudin, F., Ruigrok, R.W., Daigle, N., Ellenberg, J., Cusack, S., Simorre, J.P., Hart, D.J., 2007. Structure and nuclear import function of the C-terminal domain of influenza virus polymerase PB2 subunit. Nat. Struct. Mol. Biol. 14, 229–233.

Tsai, J., Bonneau, R., Morozov, A.V., Kuhlman, B., Rohl, C.A., Baker, D., 2003. An improved protein decoy set for testing energy functions for protein structure prediction. Proteins 53, 76–87.

Watanabe, S., Noda, T., Halfmann, P., Jasenosky, L., Kawaoka, Y., 2007. Ebola virus (EBOV) VP24 inhibits transcription and replication of the EBOV genome. J. Infect. Dis. 196 (Suppl. 2), S284–290.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., Barton, G.J., 2009. Jalview version 2 – a multiple sequence alignment editor and analysis workbench. Bioinformatics.

Westhead, D.R., Hatton, D.C., Thornton, J.M., 1998. An atlas of protein topology cartoons available on the world-wide web. Trends Biochem. Sci. 23, 35–36.

Zhang, C., Liu, S., Zhou, Y., 2004. Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. Protein Sci. 13, 391–399.

Zhou, H., Zhou, Y., 2002. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci. 11, 2714–2726.